Breakthrough in Interval Data Fitting II. From Ranges to Means and Standard Deviations

Marek W. Gutowski

Institute of Physics, Polish Academy of Sciences, 02–668 Warszawa, Poland,
email: gutow@ifpan.edu.pl

Abstract — Interval analysis, when applied to the so called problem of experimental data fitting, appears to be still in its infancy. Sometimes, partly because of the unrivaled reliability of interval methods, we do not obtain any results at all. Worse yet, if this happens, then we are left in the state of complete ignorance concerning the unknown parameters of interest. This is in sharp contrast with widespread statistical methods of data analysis. In this paper I show the connections between those two approaches: how to process experimental data rigorously, using interval methods, and present the final results either as intervals (guaranteed, rigorous results) or in a more familiar probabilistic form: as a mean value and its standard deviation.

This article is a companion paper to [1] and is meant to be its extension, but otherwise it is self-contained. This is why we don't repeat everything here, except for the most important thing: a correct way to bound the distances between uncertain experimental values and the corresponding theoretical predictions of thereof.

1 The goals of experimental data processing

The problem in front of us may be stated as follows. We have N experimental data points, labelled as $\mathbf{m}_1, \ldots, \mathbf{m}_N$ (measurements), each one obtained in different conditions $\mathbf{x}_j$, $j = 1, \ldots, N$, (called *environments* from now on), so that each $\mathbf{m}_j = \mathbf{m}_j(\mathbf{x}_j)$. In addition, we have a theory, $\mathcal{T}$, predicting the behavior of the investigated phenomenon in various environments. $\mathcal{T}$ is characterized by k $(k < N)$ unknown parameters, $\mathbf{p}_1, \ldots \mathbf{p}_k$, so formally we can write: $\mathcal{T}(\mathbf{p}_1, \ldots, \mathbf{p}_k, \mathbf{x}_j) = \mathbf{t}_j$. In words: when the (yet) unknown parameters have values $\mathbf{p}_1, \ldots, \mathbf{p}_k$ respectively, and the environment state is $\mathbf{x}_j$, the $\mathcal{T}$ predicts the observed outcome as $\mathbf{t}_j$. All quantities typeset in boldface are interval objects, usually just intervals, but they may be interval vectors as well. Contrary to the earlier theoretical attempts (for the relevant references see the literature cited in [1]) we no longer insist that experimental intervals $\mathbf{m}_j$ are guaranteed, i.e. that they contain the true values with probability equal exactly to 1, nevertheless they may have this property.

There are essentially two goals addressed by uncertain data processing:

- to determine the values of interesting parameters, $\mathbf{p}_1, \ldots \mathbf{p}_k$, best of all together with their uncertainties, or
- to test whether a given model of phenomenon under study (theory $\mathcal{T}$) is adequate.

We will not go into hypothesis testing but instead will concentrate on finding unknown parameters given the uncertain experimental information.

2 How do we find 'best fitted' parameters?

In [1] we put forward the idea that the so called 'best fits' should be based on the distance between measured and theoretical values. In one dimension, when we compare a single result of measurement with the predicted one, and at least one of those quantities is an interval, the mathematically correct distance is the one valid in the interval space $\mathbb{IR}$. Starting with the familiar Moore-Hausdorff distance [2], usually written as

$$d\left(\mathbf{a},\,\mathbf{b}\right) = \max\left(\left|\,\underline{a}-\underline{b}\,\right|,\,\left|\,\overline{a}-\overline{b}\,\right|\right), \qquad \mathbf{a},\mathbf{b}\,\in\,\mathbb{IR}, \quad d\,\in\,\mathbb{R} \tag{1}$$

we finally arrived at the tight interval estimate, $\boldsymbol{\rho}(\mathbf{t},\mathbf{m})$, of the distance between the theoretical prediction $\mathbf{t}$ and the unknown true result of a measurement, hidden somewhere within the interval $\mathbf{m}$:

- when $\mathrm{c}(\mathbf{t})\,\in\,\mathbf{m}$:

$$\begin{aligned} &\text{lower bound:} \quad \underline{\rho} \;=\; \tfrac{1}{2}\mathrm{w}(\mathbf{t}) \\ &\text{upper bound:} \quad \overline{\rho} \;=\; \max\left[\,d\left(\mathbf{t},\underline{m}\right),\,d\left(\mathbf{t},\overline{m}\right)\right] \end{aligned} \tag{2}$$

- when $\mathrm{c}(\mathbf{t})\,\notin\,\mathbf{m}$:

$$\begin{aligned} &\text{lower bound:} \quad \underline{\rho} \;=\; \min\left[\,d\left(\mathbf{t},\underline{m}\right),\,d\left(\mathbf{t},\overline{m}\right)\right] \\ &\text{upper bound:} \quad \overline{\rho} \;=\; \max\left[\,d\left(\mathbf{t},\underline{m}\right),\,d\left(\mathbf{t},\overline{m}\right)\right], \end{aligned} \tag{3}$$

where $\mathrm{c}(\cdot)$ stands for the center of its interval argument, $\mathrm{c}(\mathbf{t}) = \frac{1}{2}\left(\underline{t}+\overline{t}\right)$, and $d\left(\cdot,\cdot\right)$ is a Moore-Hausdorff distance between intervals.

Now, equipped with $\boldsymbol{\rho}$, we can think about the distances in N-dimensional spaces. They can be constructed, among other, as the counterparts of the so called L_p norms, generally defined as

$$\|\,\mathbf{x}\,\|_p = \left[\sum_{j=1}^{N}\,|x_j|^p\right]^{\frac{1}{p}}. \tag{4}$$

Here every individual x_j is a distance measured along the j-th coordinate, and p is a fixed, positive real number. The most important are norms L_1, L_2, and L_∞. Specifically we have:

- L_1 distance — a.k.a. Manhattan metric or taxi driver metric:

$$L_1(\mathbf{t},\mathbf{m}) = \sum_{j=1}^{N}\,\frac{\boldsymbol{\rho}(\mathbf{t}_j,\mathbf{m}_j)}{\mathrm{w}(\mathbf{m}_j)} \tag{5}$$

  This norm is used most often when we suspect the presence of outliers in experimental data set. The corresponding classical procedure bears the name LAD (Least Average/Absolute Deviation) optimization.

- squared L_2 norm — squared Euclidean distance:

$$L_2^2(\mathbf{t},\mathbf{m}) = \sum_{j=1}^{N} \left[\frac{\boldsymbol{\rho}(\mathbf{t}_j,\mathbf{m}_j)}{\mathrm{w}(\mathbf{m}_j)} \right]^2 \tag{6}$$

  We have shown here L_2^2 rather than just L_2, in order to underline its close relationship with familiar χ^2 functional. Minimization of χ^2, as it is well known, is an objective of the famous LSQ method. On the other hand, L_2 is a monotonous function of its positive arguments, so the minima of L_2 are located at the same arguments as minima of L_2^2.
- L_∞ or maximum distance — in interval analysis serves as box's diameter:

$$L_\infty(\mathbf{t},\mathbf{m}) = \max_{j=1,\ldots N} \frac{\boldsymbol{\rho}(\mathbf{t}_j,\mathbf{m}_j)}{\mathrm{w}(\mathbf{m}_j)} \tag{7}$$

  This metric, in turn, is best applicable for calibration purposes [3]. Here the goal is to approximate uniformly the set of experimental points via any simple curve (or surface), not necessarily physically meaningful, but easy to evaluate.

In classical data analysis every single functional shown above is treated differently than the remaining ones. Contrary, using interval methods, we need not to follow this path and develop procedures specific to each metric in turn. It is entirely possible to use exactly one and the same general purpose procedure to locate the global minimum of either functional. Such a procedure may be, for example, similar to that first described 35 years ago by Skelboe [4], and known as Moore-Skelboe algorithm.

3 Troublesome interval output

Regardless of the interval minimizer we shall use, the final outcome appears almost always troublesome. When the result is a single interval box, then the lengths of its edges are usually much larger than final uncertainties of the searched parameters as delivered by other methods. This is because such a result, being an interval hull of what was sought for, contains also many 'bad' solutions. In fact, the true solutions occupy only a little fraction of the volume returned by algorithm. Whatever the reason, our very reliable results simply look poorly, and are by no means competitive.

At the other extreme, when our minimizing algorithm delivers many boxes – and by many we mean not two, three or even dozen boxes, but rather hundreds, or maybe even thousands of them – we are in troubles again. There is no simple way to present such results to other researchers in a simple, compact form, acceptable also by publishers. Of course, we can quickly calculate the convex hull (or hulls, if the set of returned boxes is not simply connected) of all boxes, but this takes us back to the previous situation.

Even if our results happen to be quite narrow – shall we call them 'guaranteed?' Certainly not, whenever the input data have a form of the mean value and standard deviation, as it is most often the case.

Hmmm. Let's think again. Suppose, we have quite a number of boxes covering some domain in parameter space, where the true solutions are located. Aren't those points the

results of what is called 'indirect measurement?' Of course, they are! If so, then nothing can prevent us from treating them as usually and calculate their mean values, dispersion, etc.

4 Means, variances and correlations

Suppose the outcome returned by a minimizing routine is a cluster of N_{box} simply connected boxes $\mathbf{B}_j$, $j = 1, \ldots, N_{box}$, covering a single solution. How to calculate the "ordinarily" looking answers to our original problem? The number of our indirect measurements is no longer finite, as it takes place in direct measurements. It is even uncountably infinite, but this fact alone is no real obstacle. Just in place of various sums we will have to calculate some definite integrals, that's all. During calculations we have to assume that probability density is uniform in the interiors of all boxes. This position may seem strange at first sight (intervals can never be treated in this spirit!) but is entirely correct. The final formulae, valid when $N_{box} > 1$, are following:

- **mean values of unknown parameters** — center of gravity of a cluster

$$\mathbf{p}_0 = \frac{\sum_{j=1}^{N_{box}} \left[\text{center}\,(\mathbf{B}_j) \times \text{Volume}\,(\mathbf{B}_j)\right]}{\sum_{j=1}^{N_{box}} \text{Volume}\,(\mathbf{B}_j)}, \tag{8}$$

  where now $\mathbf{p}$ denotes a real-valued, k-dimensional vector of searched parameters, not their ranges: $\mathbf{p} = (p_1, \ldots, p_k)$, and the subscript '0' indicates their mean (expected) values. Of course, 'center($\mathbf{B}_j$)' is also a real-valued, k-dimensional vector, pointing – you guessed –to the center of box $\mathbf{B}_j$. The meaning of the number 'Volume($\mathbf{B}_j$)' is self-explanatory.

- **dispersions (variances) of parameters**
  We use the textbook definitions for the covariance of two multidimensional random variables $\mathbf{X}$ and $\mathbf{Y}$, when their expected values, $\mathbf{x}_0$ and $\mathbf{y}_0$, respectively, are known:

$$\text{Cov}\,(\mathbf{XY}) = \langle(\mathbf{X} - \mathbf{x}_0)(\mathbf{Y} - \mathbf{y}_0)\rangle, \tag{9}$$

  where the braces '$\langle\cdot\rangle$' mean the average (expected) value. The variance of a random multidimensional variable can be computed on two equivalent ways, either as

$$\sigma^2(\mathbf{X}) = \langle(\mathbf{X} - \mathbf{x}_0)^2\rangle \qquad \text{or as} \qquad \sigma^2(\mathbf{X}) = \text{Cov}(\mathbf{X} \cdot \mathbf{X}). \tag{10}$$

  In our case $\mathbf{X} = \mathbf{Y} = (p_1, \ldots, p_k)$. If we denote the range (interval) of parameter p_m as $\mathbf{x}$, and the range of parameter p_n as $\mathbf{y}$, both limited to current box under study, as indicated by the summation index j, then the off-diagonal elements of the covariance matrix ($m \neq n$) are expressed as:

$$\text{Cov}(p_m p_n) = \frac{\sum_{j=1}^{N_{box}} \Big[(\overline{\mathbf{x}} - 2\mathbf{x}_0)\,\overline{\mathbf{x}} - (\underline{\mathbf{x}} - 2\mathbf{x}_0)\,\underline{\mathbf{x}}\Big]_j \Big[(\overline{\mathbf{y}} - 2\mathbf{y}_0)\,\overline{\mathbf{y}} - (\underline{\mathbf{y}} - 2\mathbf{y}_0)\,\underline{\mathbf{y}}\Big]_j RV_{xy}}{4\sum_{j=1}^{N_{box}} \text{Volume}\,(\mathbf{B}_j)}, \tag{11}$$

where $\mathbf{x}_0$ and $\mathbf{y}_0$ are mean values of p_m and p_n, respectively, as computed earlier from (8). Newly introduced symbol RV_{xy} means 'reduced volume', that is the volume of $k-2$ dimensional box containing all parameters except p_m and p_n:

$$RV_{xy} = \prod_{\mathbf{Z}\neq p_m,\ \mathbf{Z}\neq p_n} (\overline{\mathbf{z}} - \underline{\mathbf{z}})_j \tag{12}$$

For diagonal elements of the covariance matrix, when $m = n$, we have instead:

$$\sigma^2(p_m) = \frac{\sum_{j=1}^{N_{box}} \left[\overline{\mathbf{x}}^2 + \underline{\mathbf{x}}\overline{\mathbf{x}} - 3\mathbf{x}_0\overline{\mathbf{x}} + \underline{\mathbf{x}}^2 - 3\mathbf{x}_0\underline{\mathbf{x}} + 3\mathbf{x}_0^2\right]_j \times \mathrm{Volume}(\mathbf{B}_j)}{3\sum_{j=1}^{N_{box}} \mathrm{Volume}(\mathbf{B}_j)} \tag{13}$$

- **correlations between parameters**
  According to any textbook on statistics, coefficient of correlation between any two multidimensional random variables is defined as:

$$\rho_{xy} = \frac{\mathrm{Cov}(\mathbf{XY})}{\sigma(\mathbf{X}) \cdot \sigma(\mathbf{Y})} \tag{14}$$

  There should be no problem with calculating this quantity when we already have all necessary ingredients, obtained from (8), (11) and (13).

5 Discussion

Omission of the case $N_{box} = 1$ was deliberate. It is both easy and hopeless case. And here is why. Easy part consists in calculating the mean values of unknown parameters. They all are simply equal to the centers of corresponding ranges. It also easy to show that their dispersions have to be equal to halves of the widths of their ranges. One will be nevertheless strongly disappointed with correlations between parameters: they are none, equal exactly to zero. But could all this be true? Certainly not.

The natural question is how accurate are the suggested here results. The boxes comprising the simply connected set covering the domain of possible solutions are not all created equal. Some of are them completely filled with the possible solutions, while others, those located at the boundaries, are filled with solutions only in part. This must necessarily affect our results, since those were derived with only the first kind of boxes in mind. It is intuitively clear that the more boxes we have, and the smaller they are, the 'filling factor' will be closer to 100 %. Consequently, our results will be closer to reality. All we can say is that the dispersions should come out always overestimated. For the cases where both input data and the theory are correct, that is. In statistical language we may say that our estimate of dispersions (or variances, if you prefer) is consistent but positively biased. Fortunately, this makes no harm.

Quite a different story concerns covariances and correlations. As we could see, our ignorance in that matter remains completely intact, when we have at our disposal only a single box. Of course, increasing the number of boxes will take us closer to the true

values. In case of off-diagonal elements of covariance/correlation matrix we have no guarantee that convergence will be one-sided. This brings us to the question how many boxes do we really need? The exhaustive answer to this problem is beyond current author's capabilities. One may hope, with analogy to other statistical problems, that sensible results should start to appear when N_{box} exceeds, say 20. Fortunately, the optimizing routine usually delivers much more boxes, counted in hundreds.

We haven't discussed the question of complexity in this paper. From what was said, it is clear that better, more accurate results, are also more costly than just rough estimates: depending on whether we are working with a single box or with many boxes.

6 Conclusions

Interval-oriented routines not only generate reliable estimates of unknown parameters as a result of uncertain data processing. So obtained results *can* be safely and reliably 'translated' into more widespread statistical form of presentation.

Interval perspective sheds completely new light on experimental data processing. Here we see with details what is in reality going on. Moreover, in many cases interval methods allow for objective estimates of accuracies, with no need for human experts (who sometimes err very much in their estimates).

Acknowledgments

This work was done as a part of author's statutory activities at the Institute of Physics, Polish Academy of Sciences.

Bibliography

[1] Marek W. Gutowski *Breakthrough in Interval Data Fitting. I. The Role of Hausdorff Distance*, in this proceedings.

[2] R.E. Moore *Interval Analysis* Prentice Hall, Englewood Cliffs, NJ, 1966

[3] A. Voschinin, N Skibitski, *Interval calibration model of multisensor system*, Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2003. Proceedings of the Second IEEE International Workshop on, pp. 253–256

[4] S. Skelboe, *Computation of rational interval functions*, BIT **14**, 87—95, 1974